METHOD AND SYSTEM FOR IDENTIFYING REPRESENTATIVE TRENDS USING SKETCHES

Reference to Related Documents

[0001] This application claims benefit of priority under 35 U.S.C. §119(e) to U.S. provisional application serial No. 60/313,473, filed August 21, 2001, the entire contents of which are incorporated herein by reference.

Technical Field

[0002] The present invention relates generally to data management and, more particularly, to methods, systems, and machine-readable media for identifying representative trends in large sets of data.

Background of the Invention

[0003] Time series databases, containing data captured over time, are commonly used in such areas as finance, meteorology, telecommunications, and manufacturing to keep track of data valuable to that particular area. For example, financial databases may track stock prices over time. Meteorological parameters such as the temperature over time are stored in scientific databases. Telecommunications and network databases include data derived from the usage of various networking resources over time such as the total number and duration of calls, number of bytes or electronic mails sent out from

one ISP to another, amount of web traffic at a site, etc.; manufacturing databases include time series data such as the sale of a specific commodity over time.

[0004] Time series data depict trends in the captured data, which users may wish to analyze and understand. Users may wish to know, for a given time window, a trend of "typical" values or an "outlier" trend. Conversely, users may wish to find the time window in which most trends are as similar as possible or clustered. These similar trends are called "representative trends." Representative trends may be used in lieu of the entire database for quick approximate reasoning. In addition, they can be used for prediction and for identifying and detecting anomalous behavior or intrusion.

[0005] By their very nature, time series databases tend to contain large amounts of data. As such, using representative trends of the data reduces the amount of data to be analyzed. However, the large amounts of data must first be processed in order to identify the representative trends.

[0006] There is a need in the art to identify representative trends efficiently and quickly in large amounts of data.

Summary of the Invention

[0007] The present invention provides a method, system, and machine-readable medium for identifying representative trends in large amounts of data using sketches. A "sketch" is a lower dimensional vector used to represent higher dimensional data. The present invention includes reducing subvectors of the data to sketches, summing the distances between each sketch and every other sketch, and selecting the data subvector corresponding to the sketch with the lowest summed distance as the representative trend of the data.

Brief Description of the Drawings

[0008] Figure 1 is a flowchart of an embodiment of a method according to the present invention;

[0009] Figure 2 is a flowchart of an exemplary method for generating sketches;

[0010] Figures 3 and 4 illustrate the method of Figure 2;

[0011] Figure 5 is a flowchart of an exemplary method for comparing sketches;

[0012] Figure 6 is a flowchart of an exemplary method for identifying representative trends in data;

[0013] Figures 7(a) – (d) illustrate representative relaxed periods and average trends; and

[0014] Figure 8 is a block diagram of an embodiment of a computer system that can implement the present invention.

Detailed Description

[0015] Embodiments of the present invention provide a method for identifying representative trends in data using sketches. A sketch is a lower dimensional vector used to represent higher dimensional data. When there are large amounts of data, the data may first be partitioned into data subvectors of a given dimension. These data subvectors can then be transformed into sketches, which have lower dimensions. The lower dimensions correspond to less data being processed. As such, using sketches of the data, rather than the data itself, provides more efficient, faster performance. So the user may be

able to quickly analyze the data without utilizing large amounts of processor time and system memory.

[0016] In addition to dimensionality reduction, sketches exhibit distance and synthesis properties that may be used in data analysis. The synthesis property provides a sketch synthesized from existing sketches. This property is particularly useful when existing sketches represent subvectors with a given dimension, but sketches are needed for subvectors with a higher dimension. In such a case, the sketch may be easily synthesized from the existing sketches, rather than calculated from scratch. This synthesis property allows the user to generate sketches once and then use those sketches to represent multiple subvector dimensions without having to reprocess the data in its entirety, thereby saving processor time and system memory.

[0017] According to the distance property, the distance between two sketches is comparable to the distance between the subvectors that the sketches represent. Thus, by calculating the distance between two sketches, the distance between the subvectors that the sketches represent may be found with measurable accuracy. In addition, by increasing the sketch dimension, the probability of identifying the data's representative trend may be increased and the error between the data and the sketches reduced. This property allows the user to calculate distances with less data – *i.e.* the sketches rather than the data itself - thereby saving processor time and system memory. The distance property holds for synthesized sketches as well.

[0018] Figure 1 is a flowchart of an embodiment of a method for identifying representative trends according to the present invention. First, sketches may be generated for data partitioned into subvectors (step 100). The data may be partitioned in a number of ways to generate the data subvectors of a given dimension T. Each adjacent subvector may include anywhere from 0 to T-1 overlapping data elements. The final subvectors may

have dimensions less than T and may generally be ignored in data analysis. The amount of overlap may affect how well synthesized sketches match their data subvectors and how quickly representative trends are identified. The sketch dimensions may be lower than the subvector dimensions. Each of the generated sketches may then be compared to every other generated sketch for a given subvector dimension T to determine how closely the sketches match each other (step 105). The sketch that has the most matches may be considered to be representative of the data. The representative trend may then be identified as the subvector corresponding to the most closely matching sketch and the period of the data may be identified as the subvector dimension T (step 110). As a result, the representative trend of the data may be found by processing less data - *i.e.,* the sketches rather than the entire data - thereby saving processing time and system memory. Optionally, this method may be repeated for multiple subvector dimensions The result may then be output to a graphical display, storage device, transmitter, or the like.

[0019] Additionally, the present invention offers the following advantages: (a) the best trend may be identified, as opposed to a sinusoidal approximation of it; (b) the trends may be identified using various metrics, as opposed to only a distance metric as is the case for Fourier transforms; (c) for noisy data, filtering may be omitted, unlike the Fourier transform; and (d) representative trends may be identified from subvectors, unlike Fourier transforms which treat the entire data.

[0020] Figure 2 is a flowchart of an exemplary method for generating the sketches. First, the sketch dimension is chosen (step 200). By choosing the sketch dimension, the user may determine how much error to allow in the final result, *i.e.* between the identified and the actual representative trend of the data. The higher the dimension, the smaller the error. Next, the data subvectors may be generated by partitioning the data (step 203). A random vector with the same dimension as the data subvectors may then be generated

to have a normalized, Gaussian distribution (steps 205, 210). The elements of the random vector may be samples from a Gaussian distribution with zero mean and unit variance, normalized to have unit length. The sketch for each data subvector may be calculated using the subvector and the random vector (step 215).

[0021] The sketch may be calculated using a dot product between the subvector and the random vector. A dot product is a well-known mathematical tool for transforming one vector into another. In the present invention, the subvector may be projected onto the random vector to produce the sketch vector. The advantage is that such a projection reduces the dimension of the data to be analyzed, thereby saving processor time and system memory.

[0022] For example, suppose $\vec{t}_1$ = (2, 1, 3, 1) and it is desired to construct a sketch vector of dimension 2. Two vectors $\vec{v}_1$ = (-0.45, -0.09, 0.10, 0.87) and $\vec{v}_2$ = (-0.19, 0.73, -0.61, 0.21) may be chosen as normalized Gaussian distributed vectors. The dot product may be calculated between $\vec{t}_1$ and $\vec{v}_1$ to produce the first element of the sketch and between $\vec{t}_1$ and $\vec{v}_2$ to produce the second element of the sketch. Hence, the sketch of $\vec{t}_1$, $S(t_1)$ is (0.18, -1.27).

[0023] Optionally, the sketch may be calculated using a polynomial convolution between the subvector and the random vector. A polynomial convolution is a well-known mathematical tool. In the present invention, the subvector may be convolved with the random vector to produce the sketch. The advantages of the convolution are that it reduces the dimension of the data to be analyzed and all the elements of the sketch vector may be calculated together, thereby saving processor time and system memory.

[0024] Figure 3 shows an example using polynomial convolution to compute sketches. A vector $\vec{t}$ = (2, 1, 3, 1) may be partitioned into subvectors

of dimension 2, t_1 = (2, 1), t_2 = (1, 3) and t_3 = (3, 1). The subvectors may then be convolved with normalized vectors $\vec{v}_1$ = (-0.97, -0.20) and $\vec{v}_2$ = (0.11, 0.99). The first and second elements of each sketch of dimension 2 may be computed at the same time, such that $S1$ = (-2.14, 1.21), $S2$ = (-1.57, 3.08), and $S3$ = (-3.1, 1.32).

[0025] Optionally, the sketch may be calculated by synthesizing it from a pool of sketches. Recall the synthesis property that allows a sketch to be synthesized from existing sketches. A pool of sketches is a small subset of the set of all sketches that could be calculated for a given set of subvectors. To generate the sketch pool, first, two sets of normalized random vectors may be generated (steps 205, 210 of Figure 2). Then, two sets of sketches may be calculated by either a dot product or a polynomial convolution using the data subvectors and each set of the random vectors. The synthesized sketch may then be calculated by adding corresponding sketches from each set. Typically, one sketch may be selected from each set. The selected sketch represents all or portions of the data to be represented by the synthesized sketch. If the dimension of the subvector of interest is a power of the subvector dimension represented in the sketch pool, then a sketch in the pool representing the same subvectors or subvector portions may be used to represent the subvector of interest. If, however, the dimension is not a power of the subvectors represented in the pool, the sketch may be synthesized as described above.

[0026] This pool of sketches may be calculated and stored prior to data analysis. As such, the pool of sketches may be used as a look-up table during analysis. Thus, the synthesized sketch may be calculated very quickly from existing sketches. This synthesis allows sketches to represent subvectors of various dimensions without recalculating random vectors and repartitioning subvectors, thereby saving processor time and system memory.

[0027] Figure 4 shows an example using a sketch pool to compute a sketch. In this example, the sketch representing a subvector of dimension 5 may be computed from a pool of sketches representing subvectors of dimension 4. The subvector of dimension 5 is $\vec{t}$ = [2 1 3 1 2]. The first set of pool sketches includes $S^1(t_1)$ = (0.09, -1.44) for t_1=[2 1 3 1] and $S^1(t_2)$ = (0.51, 1.08) for t_2 = [1 3 1 2]. The second set of pool sketches includes $S^2(t_2)$ = (0.61, 2.04) for t_2 and $S^2(t_3)$ = (0.45,0.27) for t_3 = [3 1 2 3]. The sketch pool represents subvectors having dimensions that are a power of 2. According to the present invention, since the dimension 5 is not a power of 2, the sketch for $\vec{t}$ is $S'(t)$ = $S^1(t_1) + S^2(t_2)$ = (0.70, 0.60). Note that the second, third, and fourth elements of t_1 and t_2 overlap. The more overlap between the added subvectors, the more accurate the synthesized sketch - *i.e.,* the more closely the synthesized sketch matches an actual sketch calculated from scratch. As few as one element may overlap and the accuracy may be high enough for data analysis purposes.

[0028] Figure 5 shows an exemplary method for comparing the sketches. First, sketches of subvectors of dimension T may be acquired (step 905). Then, the distance between the sketch of each subvector and the sketches of each of the other subvectors may be calculated (step 910). Exemplary distance measurements include the L_2, L_1, and L_∞ norms, which are well-known in the art. For each sketch, the inter-sketch distance may be calculated as the sum of the calculated distances (step 915). As such, a lowest of the summed distances may be found. This lowest distance indicates how closely sketches match each other and data similarities, *i.e.,* representative trends. According to the distance property, this inter-sketch distance may be substituted for the distance between the data subvectors to compare the subvectors and identify their trends. Advantageously, less data may be processed, thereby saving processor time and system memory.

[0029] Optionally, the sketch comparison may be repeated for multiple subvector dimensions T. In this case, the sketches may be recalculated or

synthesized for the different dimensions and the distances between them calculated. So, the lowest summed distance would be the lowest distance among all the sketches at all the different subvector dimensions. The advantage of employing this option is that the absolute lowest distance may be selected, indicating the best match and representative trend. This option may be used if the lowest distance exceeds a predetermined threshold, indicating that no good representative trend has been identified at the current subvector dimensions. In this instance, the data may be partitioned into subvectors of a higher dimension, T+1 for example, and the sketches generated using the pool of sketches or, optionally, from scratch.

[0030] Optionally, for each subvector dimensions T, the distance between the sketch of the first subvector and the sketches of each of the other subvectors may be calculated (step 910). For the first sketch at each T, the inter-sketch distance may be calculated as the sum of the calculated distances (step 915). This inter-sketch distance indicates how closely the first sketch matches other sketches. The lowest of the summed distances among the different dimensions may be found. This lowest distance indicates which data subvector dimension T best matches the period of the data.

[0031] After the inter-sketch distances are calculated, the representative trend may be identified and output as shown in Figure 6. The lowest inter-sketch distance may be selected (step 1000). From Figure 5, the selected distance may be the lowest distance between the first and the other subvectors among the various subvector dimensions T or the lowest distance between any one and all other subvectors among the various subvector dimensions T. The subvector dimension T that corresponds to the lowest distance may be identified as the period of the data (step 1005). As such, the subvector corresponding to the lowest distance may be identified as the representative trend of the data (step 1010). The representative trend of data may be output to a graphical display, storage device, transmitter, or the like.

[0032] The present invention may be applied to data to find relaxed periods and average trends. It is to be understood that the relaxed period and average trend applications are for exemplary purposes only, as the present invention may be used to find a variety of data patterns or trends.

[0033] A relaxed period of data t is defined as the period T of data t' generated by repeating a subvector of dimension T that most closely matches t - that is, the period T of the data t' that has the lowest distance from t. The relaxed period's representative trend is the subvector of dimension T. For example, the relaxed period's representative trend of 213123213132213 is 2132 and the relaxed period is 4. Figure 7(a) shows an exemplary data vector of dimension 15. Its corresponding trend is shown in Figure 7(b). It includes 4 repetitions of the first four values of the vector in Figure 7(a). The vector in Figure 7(b) "resembles" the original vector to a great extent. Hence the first four values of the vector in Figure 7(b) may be thought of as being representative of the entire vector of Figure 7(a).

[0034] An average trend is the subvector of data whose total distance to all the other subvectors is the smallest. The corresponding period is the subvector dimension T. For example, if t = 113123213132113 as in Figure 7(c) and T = 3, then some subvectors of interest may be 113, 123, 213, 132, and 113, or a consecutive group of three elements. The average trend is 123 which has a lowest total distance of the other subvectors. The average trend is shown in Figure 7(d). Figure 7(d) presents a vector derived by 5 repetitions of 123 in Figure 7(c). The vector in Figure 7(d) is quite similar to that in Figure 7(c), and hence may be thought of as representative. The representative trend may be output to a graphical display, storage device, transmitter, or the like.

[0035] If the distance between the sketches is zero, then the dimension of the subvectors that the sketches represent is the exact period of the data. Other variants of representative trends may be of interest as well.

[0036] Applying the method of the present invention to identify a relaxed period proceeds as follows: Data may be partitioned into subvectors of dimension T. A sketch dimension may be chosen. Then, the subvectors may be reduced to the sketches using an exemplary method, such as a dot product, polynomial convolution, or a sketch pool. If the sketch pool is used, the sketch pool would have been generated and stored prior to this process. After the sketches are generated, the distances between the first sketch and the other sketches may be calculated and summed. This may be repeated for several different subvector dimensions. Then, the lowest distance among the different dimensions may be selected. The relaxed period may be identified as the subvector dimension T corresponding to the lowest distance.

[0037] Similarly, to identify an average trend, data may be partitioned into subvectors of dimension T. A sketch dimension may be chosen. Then, the subvectors may be reduced to the sketches using an exemplary method, such as a dot product, polynomial convolution, or a sketch pool. If the sketch pool is used, the sketch pool would have been generated and stored prior to this process. After the sketches are generated, each of their distances to the other sketches may be calculated and summed for each sketch. The lowest distance may be selected. If the lowest distance exceeds a predetermined threshold, the process may be repeated for a different subvector dimension. Or the process may be repeated just to find the absolute lowest distance among several different subvector dimensions. After the lowest distance is selected, the average trend may be identified as the subvector corresponding to the lowest distance.

[0038] The methods of Figures 2, 5, and 6 may be used in combination or alternatively according to the present invention.

[0039] The present invention may be implemented for any application in which large amounts of data are used. Exemplary applications include stock market tracking and weather tracking. In such applications, a data set may be generated by sampling the measured data. For example, the price of a particular stock may be sampled every day or atmospheric pressure and temperature measurements may be sampled every hour. Conversely, the data set may be acquired from a source already in sampled form. Representative trends of the data set may then be identified. The identified trends may be output to an appropriate device for graphical display, storage, transmission, or further analysis. Exemplary analysis includes comparing the trends to prior trends to establish patterns of behavior or anomalies.

[0040] Some aspects of the present invention may be implemented using the following equations:

[0041] To synthesize a sketch, suppose there are two sketches S^1 and S^2 representing two data subvectors of dimension X, where $X < T$. The user wishes to produce a third sketch S' that represents a data subvector of dimension T. For a particular sketch - say, $S'(t[i,\ldots, i + T - 1])$ - of subvector $t[i,\ldots, i + T - 1])$, the j-th element of the sketch, where $1 \leq j \leq T$, may be synthesized as follows:

$$S'(t[i,\ldots, i + T - 1])[j] = S^1(t[i,\ldots, i+X-1])[j] + S^2(t[i + T - X,\ldots, i + T - 1])[j]). \quad (1)$$

[0042] The dimension k of a sketch may be chosen such that

$$k = \frac{9 \log L}{\varepsilon^2}, \quad (2)$$

where L is the number of subvectors of dimension T and ε is a user-defined error. By choosing k, the user also sets ε, thereby determining how much error to allow in the final result.

[0043] According to the distance property, for any given set L of subvectors of dimension T, for fixed $\varepsilon < ½$ and k, then for any pair of subvectors $\vec{t}_i, \vec{t}_j \in L$

$$(1-\varepsilon) \| \vec{t}_i - \vec{t}_j \|^2 \leq \| \vec{S}(t_i) - \vec{S}(t_j) \|^2 \leq (1+\varepsilon) \| \vec{t}_i - \vec{t}_j \|^2 . \quad (3)$$

Here $\| \vec{t}_i - \vec{t}_j \|^2$ is the L_2 distance between the two subvectors.

[0044] The distance property holds for synthesized sketches as well. In this case,

$$(1-\varepsilon) \| \vec{t}_i - \vec{t}_j \|^2 \leq \| \vec{S}'(t_i) - \vec{S}'(t_j) \|^2 \leq 2(1+\varepsilon) \| \vec{t}_i - \vec{t}_j \|^2 . \quad (4)$$

[0045] So, to compare sketches, the distance between sketches of the subvectors $\vec{S}(t_i), \vec{S}(t_j)$ may be calculated as $D(\vec{S}(t_i), \vec{S}(t_j))$, *e.g.*, using the L_2 distance. The inter-sketch distance may be calculated as the sum of the distances,

$$C^i(S(t(T))) = \sum_j D(\vec{S}(t_i), \vec{S}(t_j)) . \quad (5)$$

[0046] The mechanisms and methods of the present invention may be implemented using a general-purpose microprocessor programmed according to the teachings of the present invention. The present invention thus also includes a machine-readable medium which includes instructions which may be executed by a processor to perform a method according to the present invention. This medium may include, but is not limited to, any type of disk including floppy disk, optical disk, CD-ROMs, or any type of media suitable for storing electronic instructions.

[0047] Figure 8 is a block diagram of one embodiment of a computer system that can implement the present invention. The system 2300 may include, but is not limited to, a bus 2310 in communication with a processor 2320, a system memory module 2330, and a storage device 2340 according to embodiments of the present invention.

[0048] It is to be understood that the structure of the software used to implement the invention may take any desired form, such as a single or multiple programs.

[0049] Numerous modifications and variations of the present invention are possible in light of the above teachings. It is therefore to be understood that within the scope of the appended claims, the invention may be practiced otherwise than as specifically described herein.